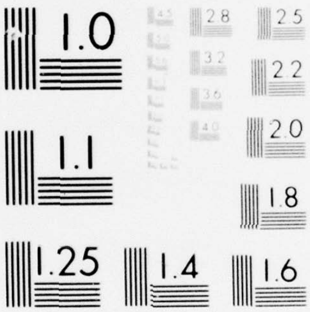

1.0
4.5
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

(11) B.S.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: (14) 11, TR-1

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): (6) DOVE, A RATIONAL ANALYSIS OF SPARSE DATA.

5. TYPE OF REPORT & PERIOD COVERED: Technical rept.
May 1, 1973-April 30, 1976

6. PERFORMING ORG. REPORT NUMBER: TR-1

7. AUTHOR(s): (10)
Peter F. Strong,
C. Gardner/Swain (principal investigator)
Marguerite S./Swain

8. CONTRACT OR GRANT NUMBER(s): (15) N00014-67-A-0204-0075

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Dept of Chemistry
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR051-566 (12) 23 p.

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research, Chemistry Program
800 North Quincy Street
Arlington, Virginia 22217

12. REPORT DATE: (11) (15) July 1[illegible] 1977

13. NUMBER OF PAGES: 21

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Distribution of this document is unlimited

(9) Technical rept. 1 May 73-30 Apr 76,

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES
~~Technical Report No. [illegible] is submitted with this report.~~
~~Both~~ will be submitted to Science (A.A.A.S.)

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
DOVE
least squares analysis
missing data
regression analysis
factor analysis

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Realistic parameters are attainable in spite of missing data. DOVE can be useful, even when many or most data are missing, for (1) generalized least squares fitting to evaluate a self-consistent set of all parameters in an expression for predicting all missing data, and (2), without changing the predicted data, to transform the set of parameters obtained in phase 1 so that each final parameter has a simple, pure, realistic, physical meaning. Since predicted data are expressed as $\underline{a}_i\underline{x}_j + \underline{b}_i\underline{y}_j + \ldots + \underline{c}_i$ with $\underline{n}$ product terms, (OVER) next page

a(i)x(j)+b(i)y(j)+...+c(i)

220 007

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered) n squared + n

cont

phase 2 requires incorporation of n^2+n independent subsidiary conditions, of which $2n$ are arbitrary, i.e., merely fix zero reference points and scale unit sizes, but n^2-n are critical, i.e., must be relationships between particular parameters supported by other information. Both phases are illustrated by a two-mode application with 7 *i*, 10 *j*, hence 41 parameters, to fit the data plus the 6 subsidiary conditions. Valid parameters are obtained although 30 of the 70 possible data are missing.

n squared − n

DOVE, a Rational Analysis of Sparse Data

Realistic parameters are attainable in spite of missing data.

Peter F. Strong, C. Gardner Swain, Marguerite S. Swain

TR-1

Mr Strong is a senior staff member at Aurthur D. Little, Inc., Cambridge, Mass. Dr. C.G. Swain is a professor of chemistry at the Massachusetts Institute of Technology, Cambridge, Mass. 02139. Dr. M.S. Swain is a postdoctoral fellow at the Massachusetts Institute of Technology.

DOVE, a Rational Analysis of Sparse Data

Realistic parameters are attainable in spite of missing data.

Peter F. Strong, C. Gardner Swain, Marguerite S. Swain

TR-1

Mr Strong is a senior staff member at Aurthur D. Little, Inc., Cambridge, Mass. Dr. C.G. Swain is a professor of chemistry at the Massachusetts Institute of Technology, Cambridge, Mass. 02139. Dr. M.S. Swain is a postdoctoral fellow at the Massachusetts Institute of Technology.

DOVE is a handy procedure for predicting missing data and forcing every parameter in the fitted expression to have a simple, discrete, realistic, physical meaning. The acronym DOVE, standing for "dual obligate vector evaluation", refers to its two-phase evaluation of all parameters, obligating them by least squares in phase 1, and additionally obligating them in phase 2 by subsidiary conditions that are supported by information other than the data (10).

Phase 1

Equation 1 embodies the least squares criterion of fit (1).

$$\sum_i \sum_j e_{ij} w_i (z_{ij} - p_{ij})^2 = \text{minimum} \qquad (1)$$

Here z and p refer to observed and predicted data, j specifies the variable of main interest, i specifies all other variables, e_{ij} is unity if z_{ij} exists but zero for any ij combination not observed, and w_i are suitable statistical weights.

Equation 2 is a generalized form of a widely applicable expression for predicted data.

$$p_{ij} = \sum_{m=1}^{n} s_{im} f_{mj} + c_i \qquad (2)$$

Its parameters comprise factors f, slopes s and intercepts c (2). However, the confusion of double subscripts on factors and slopes can be avoided by a notation using different factor and slope symbols for each different product term or mode m. Therefore we will switch to expressions for p_{ij} such as

$$p_{ij} = c_i \qquad (3)$$

$$p_{ij} = a_i x_j + c_i \qquad (4)$$

$$p_{ij} = a_i x_j + b_i y_j + c_i \qquad (5)$$

$$p_{ij} = a_i x_j + b_i y_j + g_i q_j + c_i \qquad (6)$$

or an equation with even more modes as soon as we have decided on the number, n, of modes to include.

The subscripts, j and i, need elucidation. Subscript j refers to the main or primary variable, while subscript i refers to all other variables. To be more precise, $\underline{j}$ is a numerical index for a specific example of the principal variable. In the past, this specific example has been variously called a case, individual, object, entity, or unit. Since most of these names are ambiguous or cumbersome, we will call it a "jot". Subscript $\underline{i}$ is a numerical index for a group having a common set of all the other variables. This group has also been called a variable, attribute, characteristic, property, class, or series. We will call it an "ilk" (3). For example, in a study of solvent effects the main variable is the solvent. A $\underline{j}$ of 1 might denote that water is the jot, while a $\underline{j}$ of 2 might identify the jot as ethyl alcohol. An $\underline{i}$ of 1 might refer to an ilk composed of logs of rate constants for a particular reaction at 25°C in all the solvents in which it has been studied, while $\underline{i}$=2 might mean an ilk of spectral measurements of the frequency for a particular electronic transition of a particular compound in different solvents.

Equations 4, 5 or 6 might suggest that we are only fitting a line, plane or hyperplane, respectively. If the factors (x_j, y_j, ...) were all known in advance, this would indeed be only a straightforward linear regression analysis to evaluate the $\underline{i}$-subscripted parameters. However, if any of these factors are unknown, the observed data must be used to determine $\underline{j}$-subscripted parameters as well as $\underline{i}$-subscripted parameters. Thus, in general this is a nonlinear rather than a linear least squares problem. Furthermore, the phase 1 least squares is more general than linear for another reason: any of the factors produced could prove to be a nonlinear function of one of the other factors or of several of them.

Phase 2

The least squares condition, eq. 1, is not, in general, sufficient to determine the parameters $\underline{s}_{im}$ and $\underline{f}_{mj}$ uniquely. For example, if $\underline{p}_{ij}$ satisfies eq. 4, then all the values of $\underline{a}_i$ could be doubled while all the values of $\underline{x}_j$ are halved without affecting the values of $\underline{p}_{ij}$ or the criterion of fit of eq. 1.

Therefore we propose to follow the tradition (embodied in the Brønsted catalysis law and the Hammett equation (4)) of making the factors represent conceptually simple physical influences of jots rather than only a compact means for representing or predicting data. For this purpose we usually need to transform all phase 1 parameters into new ones having a simpler and clearer interpretation, by incorporating a number of physically meaningful, independent, subsidiary conditions corroborated by other information than the data z_{ij}.

Such transformations are far from obvious when expressions as complicated as eq. 5 or 6 hold. In fact, the interpretation of observed or measured data is then always confounded and invalid conclusions about modes and parameters have usually been drawn, because the jot affects the system under study by two or more mechanisms of interaction rather than one, and the relative importance of the $\underline{n}$

mechanisms changes with both the jot and the ilk (3,5).

The total number of necessary subsidiary conditions for $\underline{n}$=2 (equation 5) is derived below, as an example. The derivation from eq. 2 for any other number of modes is similar. First express all $\underline{p}_{ij}$ (predicted data from a converged least-squares solution) as the product of a row vector $\underset{\approx}{I}'$ of the $\underline{i}$-subscripted parameters times a column vector $\underset{\approx}{J}'$ of the $\underline{j}$-subscripted parameters and unity (6). The primes indicate values calculated in phase $\underline{1}$.

$$\underset{\approx}{I}_i' = (\underline{a}_i' \quad \underline{b}_i' \quad \underline{c}_i'); \qquad \underset{\approx}{J}_j' = \begin{pmatrix} \underline{x}_j' \\ \underline{y}_j' \\ 1 \end{pmatrix}$$

$$\underline{p}_{ij} = \underset{\approx}{I}_i'\underset{\approx}{J}_j' = \underline{a}_i'\underline{x}_j' + \underline{b}_i'\underline{y}_j' + \underline{c}_i'$$

No individual $\underline{p}_{ij}$ is changed by insertion of the 3x3 unity (identity) matrix $\underset{\sim}{U}$ or its equal $\underset{\sim}{T}^{-1}\underset{\sim}{T}$ between $\underset{\approx}{I}_i'$ and $\underset{\approx}{J}_j'$.

$$\underline{p}_{ij} = \underset{\approx}{I}_i'\underset{\sim}{U}\underset{\approx}{J}_j' = \underset{\approx}{I}_i'\underset{\sim}{T}^{-1}\underset{\sim}{T}\underset{\approx}{J}_j'$$

$$\underset{\sim}{T} = \begin{pmatrix} \underline{t}_{11} & \underline{t}_{12} & \underline{t}_{13} \\ \underline{t}_{21} & \underline{t}_{22} & \underline{t}_{23} \\ 0 & 0 & 1 \end{pmatrix}$$

but the components of the resulting vectors

$$\underset{\approx}{I}_i = \underset{\approx}{I}_i'\underset{\sim}{T}^{-1} = (\underline{a}_i \quad \underline{b}_i \quad \underline{c}_i) \tag{7}$$

$$\underset{\approx}{J}_j = \underset{\sim}{T}\underset{\approx}{J}_j' = \begin{pmatrix} \underline{x}_j \\ \underline{y}_j \\ 1 \end{pmatrix} \tag{8}$$

are constants that are equally good solutions of eq. 1.

$$\underline{p}_{ij} = \underset{\approx}{I}_i\underset{\approx}{J}_j = a_i x_j + b_i y_j + c_i \tag{9}$$

Five transformation equations (10-14, representing each parameter) derived from equations 7 and 8 convert the old (primed) set of parameters to the new (unprimed) set. Matrix $\underset{\sim}{T}^{-1}$, the inverse of $\underset{\sim}{T}$, was used to derive equations 12-14.

$$\underline{x}_j = \underline{t}_{11}\underline{x}_j' + \underline{t}_{12}\underline{y}_j' + \underline{t}_{13} \tag{10}$$

$$\underline{y}_j = \underline{t}_{21}\underline{x}_j' + \underline{t}_{22}\underline{y}_j' + \underline{t}_{23} \tag{11}$$

$$\underline{a}_i = (\underline{t}_{22}\underline{a}_i' - \underline{t}_{21}\underline{b}_i')/\underline{\mathrm{det}} \tag{12}$$

$$\underline{b}_i = (\underline{t}_{11}\underline{a}_i' - \underline{t}_{12}\underline{b}_i')/\underline{\mathrm{det}} \tag{13}$$

$$\underline{c}_i = \underline{c}_i' + [(\underline{t}_{12}\underline{t}_{23} - \underline{t}_{13}\underline{t}_{22})\underline{a}_i' + (\underline{t}_{13}\underline{t}_{21} - \underline{t}_{11}\underline{t}_{23})\underline{b}_i']/\underline{\mathrm{det}} \tag{14}$$

$$\underline{\mathrm{det}} = (\underline{t}_{11}\underline{t}_{22} - \underline{t}_{12}\underline{t}_{21}) \tag{15}$$

Obviously, $\underset{\sim}{T}$ must be chosen so that det $\neq 0$.

There are six degrees of indeterminacy since six elements $\underline{t}_{11}$ to $\underline{t}_{23}$ are unspecified. To remove this indeterminacy, we must specify six independent subsidiary conditions and use them to evaluate these six elements.

Four of the six necessary conditions are trivial in this example where $\underline{n}=2$, because 2 references and 2 standards may be specified arbitrarily. Equating the factor for one jot in each mode to a reference value (commonly zero) is analogous to choosing average sea level as a height reference or the freezing point of water as a temperature reference. Equating a particular factor or slope to a standard value (never zero, commonly unity) is analogous to choosing the meter as a standard of length or K as a unit of temperature. It merely fixes the size of the scale or units in which factors for that mode are expressed.

The remaining two subsidiary conditions are critical ones and should be chosen with care and clearly stated, because they do have physical meaning and must be substantiated by other information than the data $\underline{z}_{ij}$ to ensure that all of the transformed factors and slopes will be physically simple and meaningful.

In general, the total number of necessary subsidiary conditions is $\underline{n}^2+\underline{n}$, of which $2\underline{n}$ are trivial and $\underline{n}^2-\underline{n}$ are critical.

The transformation of the parameters is required only once, after convergence has been reached in phase 1, and is in fact much simpler in program coding and much faster in computer execution time than any one of the iterative cycles preceding convergence. Nevertheless, more prior thought and more careful judgement is required in phase 2 than in phase 1.

There are circumstances where subsidiary conditions and the corresponding parameter transformation of phase 2 are unnecessary. First, one might want to know the correlation coefficient between observed $\underline{z}$ and predicted $\underline{p}$ data corresponding to one or more eq. 2 expressions for $\underline{p}_{ij}$. Neither correlation coefficients nor

$\underline{p}_{ij}$ values are changed by phase 2. The number of modes could be deduced as the $\underline{n}$ value that gives the highest correlation coefficient. Second, one might want to use one of the expressions, probably the one yielding the highest correlation coefficient, to estimate unmeasured or missing data. Although principal components and other standard factor analyses cannot be relied on when there are missing data (3), DOVE can. However, no meaning or significance can be attached to the parameters produced by phase 1 other than their ability to predict data, because they are only one set out of multiple infinities of sets, all equally good for reproducing the observed data and predicting missing data.

On the other hand, if the required number of critical subsidiary conditions can be stated and justified as true, phase 2 can be used to sort out realistically the underlying influences of different jots, and the sensitivities to these influences in different environments (ilks). These parameters can give considerably more insight into forces and mechanisms than the measured or predicted data. This is the intended purpose of factor analysis (5).

An Example Using Equation 5

DOVE was developed as an essential tool to solve the chemical problem of separating substituent effects into field and resonance components. After proving highly successful for this purpose, it was applied to separating numerous solvent effects into contributions associated with anion solvation and cation solvation. Both applications will be published separately in chemical journals. However, we anticipate that DOVE will be as or more useful in many other fields of science, engineering, and management. Since we want to prove that this procedure does yield correct answers when other methods fail, and to explain it clearly to encourage its more widespread use, we will illustrate it here by a synthetic but easily understood geometric example which we used to test the procedure because the answers are known. This is the problem of using data on 7 properties (ilks in Table 1) of 10 solid right circular cylinders (of which 3 are pictured in Figure 1) to evaluate, for each cylinder, the factors (measures or functions of radius and height) responsible for variations in the data from one cylinder to another, and to evaluate, for each property, the slopes (relative sensitivities to these factors) responsible for variations in the data from one property to another. We are pretending that we have not yet discovered a way to measure radii and heights, but wish to calculate them from measurements of these 7 other properties of the 10 cylinders. Otherwise this is a fairly realistic example for showing the kinds of limitations on such evaluations likely to arise from inability to measure underlying factors directly.

We converted the data to log data (listed in Table 2) because a DOVE phase 1 analysis on the raw data gives an overall correlation coefficient of only 0.931 with 2 modes (6 modes would be needed) but logarithms give 1.000 with 2 modes. We chose this example because this behavior is typical of several real physical problems where logarithms of measured quantities are more simply interpreted than the raw data. In chemistry, for example, one uses logs of rate constants or equilibrium constants in any attempted correlations between structure and reactivity because they are linear functions of energy differences between structures. Many sense responses (brightness, loudness, pitch) also appear to be logarithmic in character.

The input data in Table 2 could have been logs of measured data. However, instead we calculated them for cylinders having the randomly selected radii and heights (7) shown in Table 3. Now we will pretend not to know any of the formulas in Table 1 nor the 20 factors (log $\underline{r}$ and log $\underline{h}$ values) nor their 14 slopes in the logarithmic formulas but proceed to deduce them from Table 2 and subsidiary conditions only, then check all these deductions by Tables 1 and 3.

The most time-consuming phase of the analysis is the iterative adjustment of the parameters until they satisfy eq. 1. In the first half of each cycle we use multiple linear regression to calculate $\underline{a}_i$, $\underline{b}_i$, and $\underline{c}_i$ from the observed data $\underline{z}_{ij}$ and the current factor values (initially random numbers); in the next half cycle we use multiple linear regression to calculate factors from the data and the i-subscripted parameters. Further details are given under "Phase 1 Details". It is not necessary to incorporate any subsidiary conditions prior to convergence.

The slope parameters before phase 2 as we obtained them from 70 and from 50 data are shown in Table 4. They are complicated hybrid functions of the real sensitivities to radii and heights. Phase 2 unscrambles them to give simple direct measures of these sensitivities.

For the phase 2 transformation, we arbitrarily choose $\underline{x}_5=0$, $\underline{y}_5=0$, $\underline{a}_1=1$, and $\underline{b}_3=1$ as the four trivial conditions. Therefore, the factors will become differences above or below those of the reference jot, cylinder 5, taken as zero, while first-mode slopes will become ratios relative to that of the property 1 as a standard, and second-mode slopes will become ratios relative to that of property 3, another standard.

For the first of the two critical conditions we specify $\underline{b}_1=0$, reflecting possible insight that face areas of cylinders are independent of cylinder height (or of the second-term factor y_j), even though we have not yet deduced the functional form of their dependence on the first-term factor x_j nor yet determined any factors quantitively from the analysis.

For the second of the critical conditions (the last condition) we choose $a_3=a_1/2$ (or its equivalent, $a_3=\frac{1}{2}$, since $a_1=1$ is used as a trivial condition), because of three convictions: (1) that total flat face area ($i=1$) is just a multiple of the circular area of one end (although we need not even know that the multiplier λ_1 is 2); (2) that curved area ($i=3$) is proportional to circumference with a proportionality constant λ_2 that is independent of radius but is a function of height (although we need not know what function it is, i.e., that λ_2 equals height itself); (3) that circumference is proportional to the square root of curved area (although we need not know the dependence of either on radius, nor that the proportionality constant λ_3 is $2\sqrt{\pi}$). These four statements to determine critical conditions can be reasonably inferred from simple theoretical considerations or from suitable observations of another kind: they cannot be deduced uniquely from the input data of Table 2 Combining, taking logarithms, and using eq. 5 for both ilks, we obtain $a_3x_j=a_1x_j/2+\lambda_4$, where λ_4 is independent of the first factor (although it includes b_3, y_i, c_1, c_3, λ_1, λ_2, and λ_3). Since this identity must hold for all values of x_j, if follows that $\lambda_4=0$ and $a_3=a_1/2$.

Alternatively and equivalently, we could replace the last condition by $a_7=-a_1$ (or $a_7=-1$), reflecting either a theory or observations that wire cross-section and resistance are inversely related, even though we have not yet deduced a complete formula for either. Although one might expect that another alternative for the last condition could be $a_2=a_6$, implying that radius is equally influential in affecting masses or volumes, that condition is found to be ineffective for separating the factors (because the data for $i=2$ and $i=6$ also have the same dependence on height). Other undesirable assumptions are orthogonality or zero covariance between the factors because they give wrong answers in this cylinder problem, and are unlikely to be satisfied by any small sample.

Substitution of the chosen set of six conditions into transformation equations 10-13 gives

$$x_5 = 0 = t_{11}x_5' + t_{12}y_5' + t_{13}$$

$$y_5 = 0 = t_{21}x_5' + t_{22}y_5' + t_{23}$$

$$a_1 = 1 = (t_{22}a_1' - t_{21}b_1')/\underline{det}$$

$$b_3 = 1 = (t_{11}a_3' - t_{12}b_3')/\underline{det}$$

$$b_1 = 0 = (t_{11}a_1' - t_{12}b_1')/\underline{det}$$

$$\underline{a}_3 = \tfrac{1}{2} = (\underline{t}_{22}\underline{a}_3 \; - \underline{t}_{21}\underline{b}_3)/\underline{det}$$

Solution of these six simultaneous equations give the six *t* values, which may be substituted back into transformation equations 10-15 to give transformed parameters consistent with these six conditions. This transformation converts the previous parameters from either 70 or 50 data to the desired pure parameters shown in Fig. 2 and listed in the last columns of Table 4.

The slopes $\underline{a}_i$ are all exactly half of the coefficients of log *r* in Table 1. The $\underline{a}_i$ slopes (and also the $\underline{b}_i$ slopes) are thus in correct ratios relative to one another. The factor of one-half derives from one of our four less significant subsidiary conditions, a_1=1. It merely puts $\underline{a}_i$ values on a scale relative to $\underline{a}_1$ for property 1 as unity. In most applications of these numbers only relative values are needed, so it does not matter that these relative $\underline{a}_i$'s have only half of their absolute values. A slightly different way of viewing the effect of the trivial condition $\underline{a}_1$=1 is to say that it makes the $\underline{x}_j$ factors be 2 log *r* (or log $\underline{r}^2$) instead of log *r*, *i.e.*, measures of flat face area rather than of radius. This is a trivial difference because radius, diameter, circumference, and flat face area would all be equally valid quantities for the height-independent factors $\underline{x}_j$ to be representing, so the choice among them can be arbitrary.

Regardless of the choice of trivial conditions, the use of valid critical conditions lets us deduce that mass has the same dependence on radius as flat face area (since a_2=a_1), moment of inertia is twice as dependent on radius (since $\underline{a}_4$=2$\underline{a}_1$) and all the properties except 1 and 5 show the same dependence on height. Such deductions about relative factors and slopes and the functional forms of the properties are as detailed as one could expect from any kind of least squares procedure. This result obtained either without or with missing data thus seems useful and quite satisfactory.

Phase 1 Details

Least squares (1) is simply a mathematical method of fitting data, invoked because data are generally imperfect. Data that are believed to be products of various unknown powers of the factors should be linearized by taking logarithms, as illustrated in our example. Prior recognition or evidence for a linear relationship is not a prerequisite for a valid DOVE analysis, but can often simplify it by decreasing *n*.

Let the number of different ilks (i) be u, and the number of different jots (j) be v, and the number of observed data (z_{ij}) be d. Usually data are available for only a small fraction of the maximum of u times v combinations, but all that are available and believed to be reliable should be used in the analysis. To distinguish between data to be used and data that are missing or rejected, we make e_{ij} unity if the corresponding z_{ij} is to be used, otherwise zero. Although we have not used weights to reflect differences in precision or reproducibility of different data (because many measurements are made or reported only once) nor accuracy (because that is even harder to evaluate), we do use weights to make the final correlation coefficient for an ilk independent of its range. Therefore we equate each weight w_i in eq. 1 to the reciprocal of variance of the z_{ij} data for all j from the mean of z_{ij} for that ilk.

$$w_i = \frac{(\sum_j e_{ij})-1}{\sum_j e_{ij}(z_{ij}-(\sum_j e_{ij}z_{ij})/\sum_j e_{ij}\ )^2} \tag{16}$$

For eq. 5, simultaneous equations of forms 17-19 for each ilk and 20-21 for each jot

$$a_i \sum_j e_{ij}x_j^{\ 2} + b_i \sum_j e_{ij}x_j y_j + c_i \sum_j e_{ij}x_j = \sum_j e_{ij}x_j z_{ij} \tag{17}$$

$$a_i \sum_j e_{ij}x_j y_j + b_i \sum_j e_{ij}y_j^{\ 2} + c_i \sum_j e_{ij}y_j = \sum_j e_{ij}y_j z_{ij} \tag{18}$$

$$a_i \sum_j e_{ij}x_j + b_i \sum_j e_{ij}y_j + c_i \sum_j e_{ij} = \sum_j e_{ij}z_{ij} \tag{19}$$

$$x_j \sum_i e_{ij}w_i a_i^{\ 2} + y_j \sum_i e_{ij}w_i a_i b_i = \sum_i e_{ij}w_i a_i(z_{ij}-c_i) \tag{20}$$

$$x_j \sum_i e_{ij}w_i a_i b_i + y_j \sum_i e_{ij}w_i b_i^{\ 2} = \sum_i e_{ij}w_i b_i(z_{ij}-c_i) \tag{21}$$

are obtained by substituting p_{ij} into eq. 1 and then setting the $3u + 2v$ partial derivatives of eq. 1 with respect to each parameter equal to zero. Beginning with random numbers for x_j and y_j values, one uses the u sets of three equations (17-19) in three unknowns to solve for values of $\underline{a}_i$, $\underline{b}_i$ and $\underline{c}_i$. These are then used in the $\underline{v}$ sets of two equations (20-21) in two unknowns to solve for better values of $\underline{x}_j$ and $\underline{y}_j$. Thus by the successive approximation method of solving these equations alternately, one of the infinite number of sets of values for the $3\underline{u} + 2\underline{v}$ constants consistent with eq. 1 and 5 is finally obtained. This particular converged set, dependent on the initial random numbers, is then transformed in phase 2 into the unique set consistent with the desired six subsidiary conditions.

All calculations involving real numbers were done to a precision of 16 decimal places, using a FORTRAN IV program on an IBM 370-168 computer. While conforming logically to the above description, our program obviated storage of both missing data and the existence matrix by use of three simple arrays for measured data, $\underline{i}$, and $\underline{j}$, each singly subscripted by only a measured datum number ($\underline{k}$=1,2,...,$\underline{d}$), and by searches, when needed, through these arrays.

Convergence was achieved in a smaller number of iterative cycles by appropriate use of overrelaxation of all the $\underline{i}$-subscripted parameters (e.g., making changes in them larger than calculated by multipliers of 1.6 or more in most cycles) and by less frequent but longer extrapolations of all the factors (e.g., changing each by a common large multiple of its total change in the last one or two cycles, every 15-30 cycles). Such techniques are generally required for a practical solution. The square of the correlation coefficient, deco,

$$\text{deco} = 1-(\sum_i \sum_j \underline{e}_{ij}\underline{w}_i(\underline{z}_{ij}-\underline{p}_{ij})^2/\underline{d}(+6-3\underline{u} + 2\underline{v})$$

was calculated just before each extrapolation and two cycles later, and constancy to 12 decimal places used as a criterion of convergence. Any extrapolation resulting in a decreased deco two cycles later was effectively erased by return to the parameters existing just prior to extrapolation. Deco corrects for sample size (through degrees of freedom) and for dissimilar data ranges or unit sizes in different ilks (through weights as described above). After convergence, it represents the fraction of the variation in $\underline{z}_{ij}$ attributable to variations in the parameters explicitly included in the expression chosen for $\underline{p}_{ij}$, as opposed to errors or unidentified factors.

Many more cycles are needed if a large fraction of the possible data are missing. For example, in this cylinder problem we reached convergence in 1 cycle when 70 data were used, within 25 cycles when 50 data were used (deleting the 20 indicated by stars in Table 2), but only by 321, 350, and 417 cycles when 42, 41 and 40 data were used (8). In the last case, the data are fewer than the number of parameters determined (21 + 20 = 41), but the subsequent transformation adds 6 subsidiary conditions to the 40 data, thereby providing sufficient information to make all the final parameters unique and meaningful.

Non-linear least squares based on the Marquardt algorithm (9) is an alternative method for fitting all the parameters in these or any equations, linear or otherwise. Unfortunately, computer execution time for each cycle is extremely long by the Marquardt procedure when the number of parameters exceeds 10, being proportional to the cube of the total number of parameters, in striking contrast to DOVE, where it is proportional to the first power. In the present example with 41 parameters, total execution time for convergence is 7 sec. with 70 data and 51 sec. with 69 data (only one missing datum), vs. DOVE execution times of 0.44 sec. with 70 data, 1.4 sec. with 50 data, and 12 sec. with 40 data (30 missing data).

Phase 2 Details

Normalization conditions for the factors ($\sum_j x_j^2 = \underline{v}$ and $\sum_j y_j^2 = \underline{v}$) are less convenient for trivial conditions than selection of values for particular jots, because they change the sizes of the units in which parameters are expressed every time data are added or deleted.

For any set of subsidiary conditions to be acceptable, the determinant of T must be nonzero. In the course of deriving transformation equations for more than 100 different sets of six subsidiary conditions (appropriate to eq. 5 with different kinds of problems), we have found it always wise to check this early in the derivation. With many possible sets of subsidiary conditions, it is much easier to do two or more successive transformations, incorporating some but not all of the desired conditions in the first transformation. Substitution of the values of previously fixed parameters often considerably simplifies the derivation of equations for subsequent transformations.

DOVE has an enormous potential in many fields for correctly interpreting data expressible by equation 5, and a potential unmatched by any other method when some of the data are missing. Phase 1 should also be applicable to problems involving three or more modes. However, with three or more modes, the phase 2 problem of finding, justifying and incorporating the required large number of critical conditions becomes a major obstacle in DOVE or any linear least squares, nonlinear least squares, principal components or other factor analysis procedure purporting to provide meaningful parameters. Orthogonalization between factors of different modes has often been used, but the number of such conditions is insufficient, orthogonality is never satisfied by real data from statistically small samples, and often would not be satisfied even by infinite-size samples when the factors have real physical significance. For example, for substituent effects in chemistry, the resonance and other electronic (field and inductive) factors associated with substituents certainly have a small but undoubtedly physically meaningful positive correlation; and for solvent effects, the two types of factors associated with the solvent (anion-stabilizing ability and cation-stabilizing ability) clearly have a weak but significant negative correlation. To assume that they do not would force the derived factors to take on numerical values that are complicated hybrids rather than pure measures of these physical characteristics. Unless the proper number of meaningful and valid critical conditions is incorporated, the factors and slopes have no simple interpretation or meaning, even though all predicted data p_{ij} may agree very accurately with observed data z_{ij}.

Summary

DOVE can be useful, even when many or most data are missing, for (1) generalized least squares fitting to evaluate a self-consistent set of all parameters in an expression for predicting all missing data, and (2), without changing the predicted data, to transform the set of parameters obtained in phase *1* so that each final parameter has a simple, pure, realistic, physical meaning. Since predicted data are expressed as $a_i x_j + b_i y_j + \ldots + c_i$ with n product terms, phase 2 requires incorporation of $n^2 + n$ independent subsidiary conditions, of which $2n$ are arbitrary, *i.e.*, merely fix zero reference points and scale unit sizes, but $n^2 - n$ are critical, *i.e.*, must be relationships between particular parameters supported by other information. Both phases are illustrated by a two-mode application with 7 i, 10 j, hence 41 parameters, to fit the data plus the 6 subsidiary conditions. Valid parameters are obtained although 30 of the 70 possible data are missing.

References and Notes

1. A.M. Legendre, Nouvelles Methodes pour la Determination des Orbites des Cometes (Paris, 1805, pp. 72-80); M. Merriman, Trans. Conn. Acad., 4, 151 (1877).

2. An equation with an i-dependent intercept is usually desirable, for at least two reasons: (a) it allows equal statistical weighting of all z_{ij} including any for which $x_j=y_j=0$; and (b) it permits inclusion of ilks for which no z_{ij} exists for which $x_j=y_j=0$.

3. C.G. Swain, H.E. Bryndza, M.S. Swain, following article, gives references to methods and applications of standard or conventional factor analyses. Our definitions differ from those prevalent in standard factor analysis, where our main variable (the jot) is usually called a "case" and not considered a variable at all, but where an attribute associated with a particular fixed set of other variables (an ilk) is called a "variable". One should keep these distinctions in mind when reading the literature or even the definitions of factor analysis (5). Furthermore, our "factors" have often been called "factor scores", while our "slopes" have usually been called "factor loadings" or "factor coefficients". Our slopes and modes have also frequently been called factors.

4. J.E. Leffler and E. Grunwald, Rates and Equilibria of Organic Reactions (Wiley, New York, N.Y., 1963, pp. 235-241, 172-185).

5. An undercurrent of uneasy feeling about the effectiveness of available factor analyses is reflected in the definition of factor analysis as "the use of one of several methods for reducing a set of variables to a lesser number of new variables each of which is a function of one or more of the original variables [Dictionary of the English Language, Random House, New York, N.Y., 1967]. With this paper we return to the earlier, broader and sounder definition in Webster's Third International Dictionary [Merriam, Springfield, Mass. 1961] as "a statistical method for the identification of each of several variables that fluctuate together and for the determination of their relative contribution to a mingled influence", which implies a search for the pure or physically meaningful underlying influences or factors (and their quantitative evaluation) rather than the mere mathematical construction of any arbitrary or artificial set of hybrids of the original variables.

6. DOVE is "dual" in another sense: in general, for each eq. 2-6, it evaluates two vectors $\underset{\approx}{I}_i$ and $\underset{\approx}{J}_j$. Vectors are symbolized by boldface italic capital letters, matrices by boldface upright capitals, and components of vectors or matrices, or scalar quantities, by lowercase letters.

7. Subroutine RANDU of SSP [see *3*, reference 16] was used with initial integer 817799683 to generate these values.

8. We are grateful to Dr. Niles R. Rosenquist for coding and using early versions of this computer program to study chemical substituent effects, where 45 cycles were sufficient for convergence (with 220 data, 63% missing, 14 ilks, 43 jots). We thank Mr. John D. Arenivar for testing several subsets of the cylinder data and many modifications of the computer program to speed up convergence. In principle, the number of data in this cylinder problem could be as low as 35.

9. D.W. Marquardt, *J. Soc. Industrial Applied Math.*, **11**, 431 (1963).

10. This work was supported by research grants from the Office of Naval Research and the Petroleum Research Fund of the American Chemical Society.

Table 1. Cylinder properties used.

Property observed		Formula to be deduced	
$\underline{i}$	ilk	nonlinear	linear log form
1	total area of flat faces	$2\pi \underline{r}_j^{\,2}$	$2 \log \underline{r}_j + \log(2\pi)$
2	mass; $\delta = 10$ g/cm^3	$\delta\pi \underline{r}_j^{\,2} \underline{h}_j$	$2 \log \underline{r}_j + \log \underline{h}_j + \log(\delta\pi)$
3	area of curved surface	$2\pi \underline{r}_j \underline{h}_j$	$\log \underline{r}_j + \log \underline{h}_j + \log(2\pi)$
4	axle moment of inertia	$\delta\pi \underline{r}_j^{\,4} \underline{h}_j/2$	$4 \log \underline{r}_j + \log \underline{h}_j + \log(\delta\pi/2)$
5	aspect ratio	$\underline{r}_j/\underline{h}_j$	$\log \underline{r}_j - \log \underline{h}_j$
6	volume of circumscribed square prism	$4\underline{r}_j^{\,2}\underline{h}_j$	$2 \log \underline{r}_j + \log \underline{h}_j + \log 4$
7	resistance between faces; ρ=0.1 ohm cm	$\rho \underline{h}_j/\pi \underline{r}_j^{\,2}$	$-2 \log \underline{r}_j + \log \underline{h}_j + \log(\rho/\pi)$

Table 3. A group of random numbers, used to generate Table 2.

Cylinder, $\underline{j}$	Radius, $\underline{r}_j$	Height, $\underline{h}_j$	$\log(\underline{r}_j/\underline{r}_5)$	$\log(\underline{h}_j/\underline{h}_5)$
1	0.021658190	0.408169508	-1.61	-0.36
2	0.543617487	0.456423521	-0.21	-0.31
3	0.030803025	0.153893471	-1.46	-0.78
4	0.521428823	0.799775958	-0.23	-0.07
5	0.890675187	0.930589199	(0.00)	(0.00)
6	0.796412110	0.968748212	-0.05	0.02
7	0.190720081	0.059738331	-0.67	-1.19
8	0.923573375	0.606675744	0.02	-0.19
9	0.175991654	0.081358790	-0.70	-1.06
10	0.358400822	0.323721051	-0.40	-0.46

Table 2. Input data set used to test various procedures.

Jot	Ilk						
j	i = 1	2	3	4	5	6	7
1	-2.5305758*	-2.2207653	-1.255357	-5.8505510	-1.2752184*	-3.1158551	1.4424464
2	0.2687667	0.6271047	0.1928413*	-0.2033384	0.0759254	-0.2679852*	-1.3083687
3	-2.2246334†	-2.3384432	-1.5260066	-5.6622865†	-0.6986268†	-3.2335331	0.7128836†
4	0.2325699*	0.8345083*	0.4183432*	-0.0321316	-0.1857733*	-0.0605816	-1.0285716
5	0.6976186	1.3653466	0.7166572	0.9637553‡	-0.0190387†	0.4702567	-1.4278306*
6	0.6004556	1.2856365*	0.6855286*	0.7868822	-0.0850731†	0.3905466	-1.3132147
7	-0.6410273	-1.1658042§	-1.1451706	-2.9060413*	0.5041433	-2.0608941*	-1.2816896
8	0.7291227†	1.2110493	0.5466079	0.8409621	0.1825148	0.3159594*	-1.6451360*
9	-0.7108360*	-1.1014615	-1.0459236	-2.9115073	0.3350876†	-1.9965514	-1.0777295*
10	-0.0930821*	0.1160588	-0.1372802*	-1.0762332*	0.0441981	-0.7790311	-1.0957169

* One of 20 data later deleted to test the effect of missing data.

† Additional data deleted for tests with ≥ 28 missing data.

‡ Additional datum deleted for tests with ≥ 29 missing data.

§ Additional datum deleted for test with 30 missing data.

Table 4. Slopes, $\underline{a}_i$ and $\underline{b}_i$

	Before Phase 2*				After Phase 2†	
Ilk	70 data§		50 data§		70 or 50 data§	
$\underline{i}$	$\underline{a}_i$	$\underline{b}_i$	$\underline{a}_i$	$\underline{b}_i$	$\underline{a}_i$	$\underline{b}_i$
1	1.66	2.30	1.98	3.01	(1.00)‡	(0.00)‡
2	1.19	3.19	1.96	4.40	1.00	1.00
3	0.36	2.04	0.97	2.89	(0.50)‡	(1.00)‡
4	2.85	5.49	3.93	7.41	2.00	1.00
5	1.31	0.26	1.01	0.12	0.50	-1.00
6	1.19	3.19	1.96	4.40	1.00	1.00
7	-2.14	-1.42	-1.99	-1.63	-1.00	1.00

* After convergence to meet the least squares conditions but before parameter transformations to incorporate six subsidiary conditions.

† Relative values after incorporation of six subsidiary conditions.

‡ Value specified by one of the six subsidiary conditions.

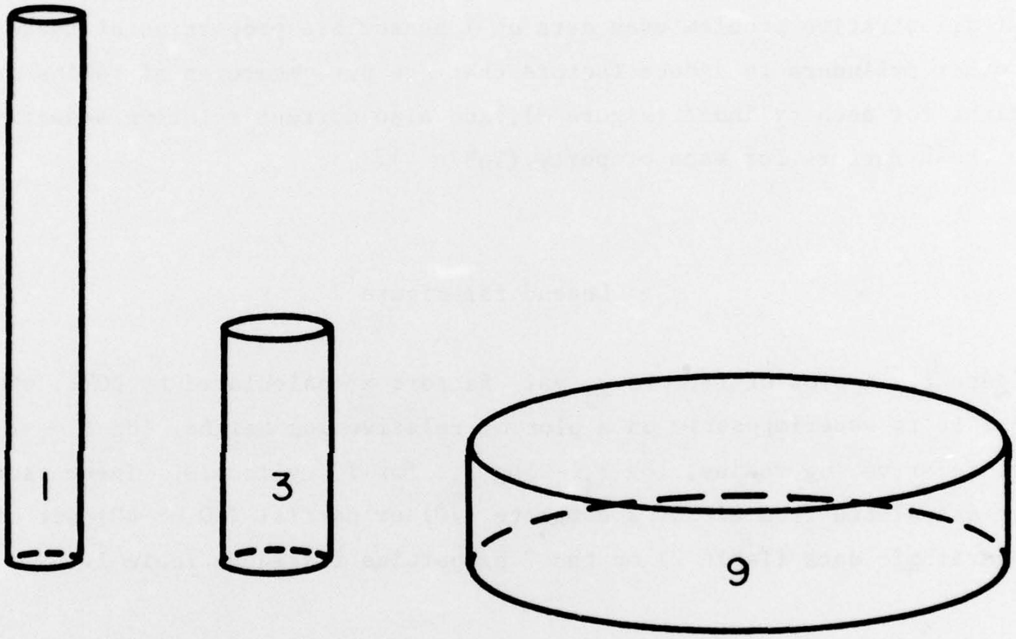
§ Number of input data $\underline{z}_{ij}$ used in the analysis.

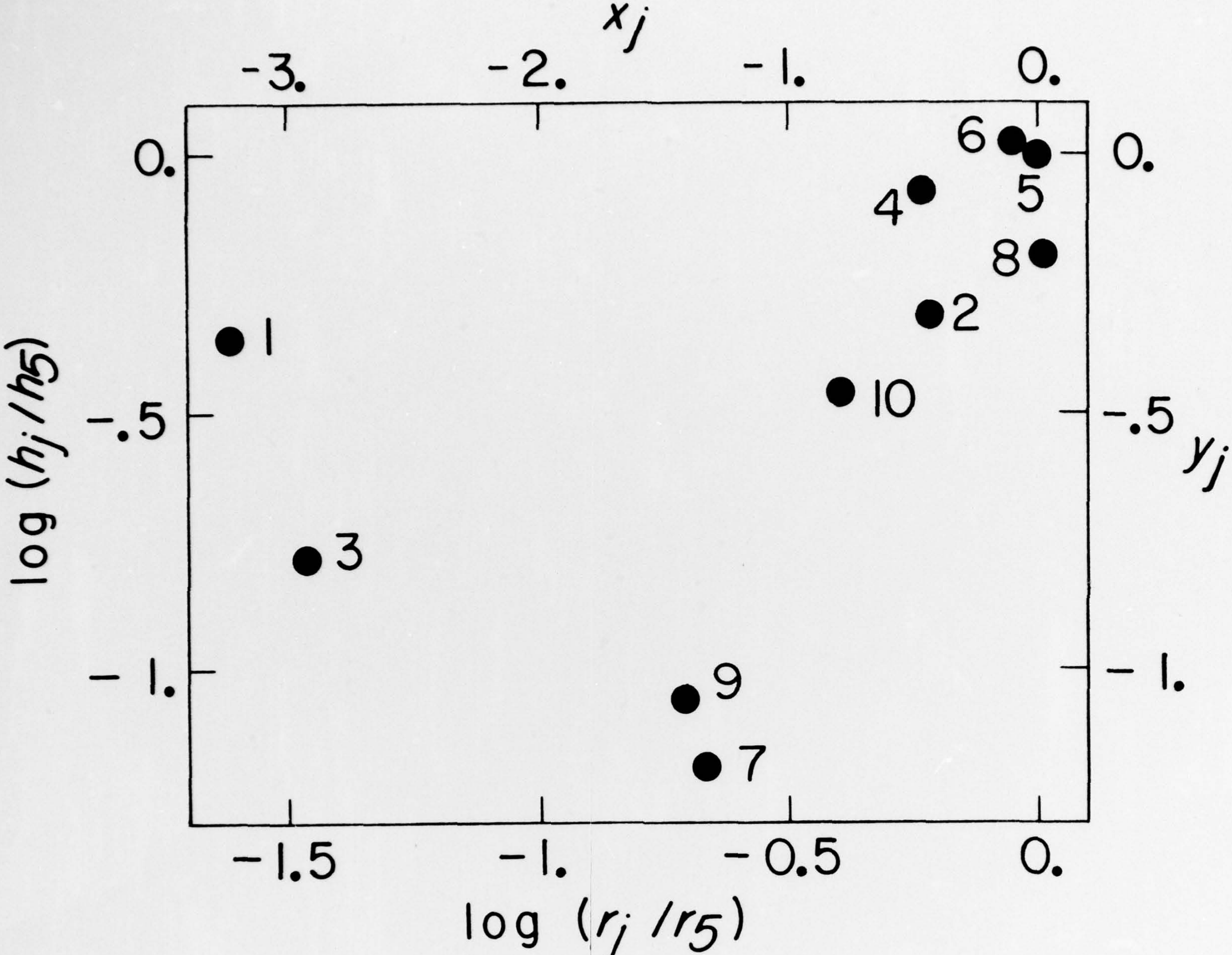
Legend for Figure 1

Figure 1. Shapes of cylinders 1, 3, and 9 in the example. This synthetic but illustrative problem uses data on 7 measurable properties of these and 7 other cylinders to deduce factors that are pure measures of radius or height for each cylinder (Figure 2), and also correct relative sensitivities to these factors for each property (Table 4).

Legend for Figure 2

Figure 2. A plot of factors $\underline{y}_j$ _vs_. factors $\underline{x}_j$ calculated by DOVE, showing that it is superimposable on a plot of relative log height, log $\underline{h}_j$ - log $\underline{h}_5$, _vs_. relative log radius, log $\underline{r}_j$ - log $\underline{r}_5$, for 10 cylinders. These factors are calculated from either a complete (70) or partial (50 or 40) set of logarithmic data (Table 2) on the 7 properties listed in Table 1.


1
3
9



x_j
-3.
-2.
-1.
0.
0.
-.5
-1.
log (h_j/h_5)
0.
-.5
y_j
-1.
1
2
3
4
5
6
7
8
9
10
-1.5
-1.
-0.5
0.
log (r_j/r_5)